Claims

1. A method of searching content, in which at least one extract common to a first file and to a second file, in the form of binary data, is searched for, characterized in that it comprises a prior preparation of the first file at least, comprising the following steps:

a) segmenting the first file into a succession of data packets, of chosen size, and identifying addresses of packets in said file,

b) associating with the address of each packet a digital signature defining a fuzzy logic state from among at least three states: "true", "false" and "undetermined", said signature resulting from a combinatorial calculation on data emanating from said file,

and in that the method continues with a search for common extract, itself, comprising the following steps:

c) comparing the fuzzy logic states associated with each packet address of the first file, with fuzzy logic states determined on the basis of data emanating from the second file,

d) eliminating from said search for common extract, pairs of respective addresses of the first and second files whose respective logic states are "true" and "false" or "false" and "true", and preserving the other pairs of addresses identifying data packets liable to comprise said common extract.

2. The method as claimed in claim 1, characterized in that, in step b), a data packet is assigned the state:

- "true" if all the data of the packet satisfy a first condition,
- "false" if all the data of the packet satisfy a second condition, contrary to the first condition,
- and "undetermined" if certain data of the packet

satisfy the first condition, while other data of the packet satisfy the second condition.

3. The method as claimed in one of claims 1 and 2, characterized in that a processing prior to step b) is applied to the data of a file, said processing comprising the following steps:

a1) the data of the file are considered as a string of samples obtained at a predetermined sampling frequency (Fe), and of values coded according to a binary representation code, and

a2) a digital filter is applied to said samples (f_n), said filter being adapted to minimize a probability of obtaining the "undetermined" state for the digital signatures associated with the packets of samples.

4. The method as claimed in claim 3, characterized in that the application of said digital filter amounts to:

- applying a spectral transform to the sampled data,
- applying a low-pass filter to said spectral transform,
- and applying an inverse spectral transform after said low-pass filter.

5. The method as claimed in claim 4, characterized in that the low-pass filter operates on a frequency band comprising substantially the interval:

[-Fe/2(k-1), +Fe/2(k-1)],

where Fe is said sampling frequency,
and k is the number of samples that a packet comprises.

6. The method as claimed in one of claims 4 and 5, characterized in that said digital filter comprises a predetermined number of coefficients of like value, and in that the frequency response of the associated low-pass filter is expressed, as a function of frequency f, by an expression of the type:

sin(PI.f.T)/(PI.f.T),

where sin() is the sine function, and with:

- PI = 3.1416, and
- T=(K-1)/Fe where K is said predetermined number of coefficients and Fe said sampling frequency.

7. The method as claimed in one of claims 3 to 6, characterized in that said digital filter is a mean value filter of a predetermined number (2K+1) of coefficients, and in that the difference between two successive filtered samples ($r_{n+1}-r_n$) is proportional to the difference between two unfiltered samples ($f_{n+K+1}-f_{n-K}$), respectively of a first rank and of a second rank, which are spaced apart by said predetermined number of coefficients, and in that the calculation of said filtered samples is performed by utilizing this relation to reduce the number of calculation operations to be performed.

8. The method as claimed in one of claims 6 and 7, characterized in that said predetermined number of coefficients of the filter (2K+1) is greater than or equal to 2k-1, where k is the number of samples that a packet comprises.

9. The method as claimed in one of claims 3 to 8, taken in combination with claim 2, characterized in that:

- the "true" state is assigned to the address of a packet if, for this packet, all the filtered samples have a value greater than a chosen reference value (Vref),
- the "false" state is assigned to the address of a packet if, for this packet, all the filtered samples have a value less than a chosen reference value (Vref), and
- the "undetermined" state is assigned to the address of a packet if, for this packet, the

filtered samples have, for certain of them, a value less than said reference value (Vref), and, for other filtered samples, a value greater than said reference value (Vref).

10. The method as claimed in claim 9, characterized in that, for any filtered sample r_n of given order n, said reference value (Vref) is calculated by averaging the values of the unfiltered samples f_k over a chosen number of unfiltered consecutive samples (Kref) about an unfiltered sample f_n of the same given order n.

11. The method as claimed in claim 10, characterized in that the values of the filtered samples are made relative, for comparison, to a zero threshold value, and in that said filtered samples r'_n are expressed by a sum of the type:

$$r'_n = K_{ref} \sum_{k=-(K/2)}^{(K/2)-1} f_{n+k} - K \sum_{k=-(K_{ref}/2)}^{(K_{ref}/2)-1} f_{n+k}, \text{ where:}$$

- f_{n+k} are unfiltered samples obtained in step a1),
- K is the number of coefficients of the digital filter, preferably chosen to be even, and
- K_{ref} is said number of unfiltered samples around an unfiltered sample f_n, preferably chosen to be even and greater than said number of coefficients K.

12. The method as claimed in claim 11, characterized in that said sum is applied to the unfiltered samples f_n a plurality of times, according to a processing performed in parallel, while respectively varying the number of coefficients K.

13. The method as claimed in one of the preceding claims, characterized in that the fuzzy states associated with the first file at least are each coded on at least two bits.

14. The method as claimed in claim 13, taken in combination with claim 12, characterized in that the fuzzy states determined for a least number of coefficients K are coded on least significant bits and the fuzzy states determined for a larger number of coefficients K are coded on subsequent bits, up to a chosen total number of bits.

15. The method as claimed in one of claims 3 and 10, characterized in that each filtered sample r_n is expressed as a sum of the type:

$$r_n = \sum_{i=-I_1}^{I_2} filter_i \times f_{(n+i)}$$, where:

- $f_{(n+i)}$ are unfiltered samples,
- $filter_i$ are coefficients of a digital filter, integrating, as the case may be, a threshold value referred to zero,

and in that a number k of unfiltered samples that a packet comprises is chosen, at minimum equal to 2 and less than or equal to an expression of the type: $(TEF-I_1-I_2+1)/2$, where TEF is a desired minimum size of the common extracts searched for.

16. The method as claimed in claim 15, characterized in that:

- for a given value TEF of the desired minimum size of common extracts searched for, a span of usable values for said number k of unfiltered samples that a packet comprises is determined,
- and, for each usable value of the number k, an optimal size TES is determined of a succession of data of digital signatures, for which succession the detection of a common extract of size TEF is guaranteed,

and in that said optimal size TES is less than or equal to an expression of the type:
$E[(TEF-I_1-I_2+1)/k]-1$, where E(X) designates the integer part of X.

17. The method as claimed in one of the preceding claims, in which the two files to be compared comprise data representative of alphanumeric characters, in particular of the text and/or a computer or genetic code,
characterized in that the method comprises:
- a first group of steps comprising the formation of the digital signatures and their comparison, for a coarse search, and
- a second group of steps comprising an identicalness comparison in the spans of addresses satisfying the coarse comparison,

and in that the data of a file are considered as a string of samples, with a chosen number k of samples per packet,
and in that the value of this chosen number k is optimized initially by searching for a minimum of comparison operations to be performed.

18. The method as claimed in claim 17, characterized in that, for the optimization of the chosen number k of samples per packet, account is taken of a total number:
- of operations of comparison of digital signatures to be performed, and
- of operations of identicalness comparison of data to be performed thereafter,

and in that said total number of operations is a minimum for a finite set of numbers k.

19. The method as claimed in one of claims 17 and 18, characterized in that an information relating to a minimum desired size of common extracts searched for (TEF) is obtained, used to optimize said chosen

number k of samples per packet,
and in that the optimal number k of samples per packet varies substantially as said minimum size (TEF), so that the larger desired minimum size of common extracts searched for, the shorter the duration of the search for common extract.

20. The method as claimed in one of claims 1 to 16, characterized in that it comprises the search for common extracts consists of a single group of steps comprising the formation of the digital signatures and their comparison, and in that the number of data items per packet is optimized by initially fixing a confidence index characterizing an acceptable threshold of probability of false detection of common extracts.

21. The method as claimed in one of claims 3 to 20, characterized in that, for the first file:

- the sampling at a chosen sampling frequency,
- the digital filtering corresponding to a low-pass filtering in the frequency space, and
- the combination of the filtered samples to obtain digital signatures in the "true", "false" or "undetermined" state, associated with the respective addresses of the first file,

while it comprises, for the second file:

- the sampling at a chosen sampling frequency,
- the digital filtering corresponding to a low-pass filtering in the frequency space, and
- the logic state associated with each packet of filtered samples is determined on the basis of the logic state associated with a single filtered sample chosen from each packet,

in such a way as to obtain digital signatures comprising only "true" or "false" logic states and thus to improve the selectivity of the comparison of the digital signatures.

22. The method as claimed in claim 21, characterized in that,

- if the logic state associated with an address of the first file is "true" or "undetermined", while the logic state associated with an address of the second file is "true", the pair of said addresses is retained from the search of common extract,
- if the logic state associated with an address of the first file is "false" or "undetermined", while the logic state associated with an address of the second file is "false", the pair of said addresses is retained for the search for common extract,

while the other pairs of addresses are excluded from the search.

23. The method as claimed in claim 20, in which the first and second files are files of samples of digitized signals, characterized in that the method comprises a step of preprocessing of the data and a taking into account of the data associated with portions of signal of level greater than a noise reference.

24. The method as claimed in one of claims 20 and 23, in which the first and second files are files of samples of digitized signals, characterized in that the method provides for a step of consolidation of the search results, preferably by adjustment of relative sizes of the packets of the first and second files, in such a way as to tolerate a discrepancy in respective speeds of retrieval of the first and second files.

25. The method as claimed in one of the preceding claims, characterized in that one at least of the first and second files is a data stream,
and in that the method of searching for common extracts is executed in real time.

26. A computer program product, intended to be stored in a memory of a central unit of a computer or on a removable medium intended to cooperate with a reader of said central unit, characterized in that it comprises instructions for conducting all or part of the steps of the method according to one of the preceding claims.

27. A data structure intended to be used for a search of at least one extract common to a first and a second file, the data structure being representative of the first file,
characterized in that it is obtained by the implementation of steps a) and b) of the method as claimed in one of claims 1 to 25,
and in that it comprises a succession of addresses identifying addresses of the first file and to each of which is assigned a fuzzy logic state from among the states: "true", "false" and "undetermined".

28. A computer device, comprising a memory for storing at least first and second files, for the search for at least one extract common to the first file and the second file, characterized in that it comprises a memory suitable for storing the instructions of a computer program product as claimed in claim 26.

29. A computer installation, comprising:
- a first computer entity suitable for storing a first file,
- a second computer entity suitable for storing a second file, and
- means of communications between the first and second computer units,

characterized in that one of the entities at least comprises a memory suitable for storing the computer program product as claimed in claim 26, for the search of extract common to the first and second files.

30. The installation as claimed in claim 29, characterized in that the entity storing the computer program product is designed to perform a remote update of one of the first and second files with respect to the other of the first and second files.